Методы решения сеточных уравнений

1 Прямые и итерационные методы

В результате разностной аппроксимации краевых и начально-краевых задач математической физики получаются СЛАУ, матрицы которых обладают следующими свойствами:

1) порядок матрицы очень высок и равен числу узлов сетки;

2) матрицы являются разреженными и имеют большое число нулевых элементов;

3) матрицы являются плохо обусловленными: отношение наибольшего собственного значения матрицы к ее наименьшему собственному значению является величиной $O(|h|^{-2})$.

Методы решения СЛАУ, получаемых в результате разностной аппроксимации исходной задачи, подразделяют на прямые и итерационные.

Прямыми методами решения задач называют методы, с помощью которых можно получить точное решение задачи за конечное число арифметических операций.

Итерационными называют методы, с помощью которых можно получить приближенное решение задачи с любой заданной точностью за конечное число арифметических операций.

2 Прямые методы

2.1 Метод монотонной прогонки и метод окаймления

Основным прямым методом, используемым для решения СЛАУ, возникающих в разностных схемах для начально-краевых задач математической физики, является метод прогонки и различные его варианты. Классический вариант метода прогонки, называемый также монотонной правой прогонкой, представляет собой метод Гаусса без выбора главного элемента для систем с трехдиагональной матрицей. Напомним, что для системы

вида

$$\begin{cases} -c_0 y_0 + b_0 y_1 = -f_0, \quad (i = 0), \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \quad (i = 1, 2, ..., N-1), \\ a_N y_{N-1} - c_N y_N = -f_N, \quad (i = N) \end{cases} \tag{2.1}$$

метод правой монотонной прогонки позволяет получить систему

$$\begin{cases} y_N = \beta_N, \\ y_i = \alpha_i y_{i+1} + \beta_i, \ \ i = N-1, N-2, ..., 0, \end{cases}$$

где

$$\alpha_0 = \frac{b_0}{c_0}; \ \ \beta_0 = \frac{f_0}{c_0}; \ \ \alpha_i = \frac{b_i}{c_i - a_i \alpha_{i-1}}, \ i = 0, 1, ..., N-1; \quad \beta_i = \frac{f_i + a_i \beta_{i-1}}{c_i - a_i \alpha_{i-1}}, \ i = 0, 1, ..., N.$$

Для решения системы (2.1) данным методом необходимо затратить $Q = 8N + 1$ действий. Метод правой монотонной прогонки корректен, если выражения в знаменателе для коэффициентов α_i, β_i не обращаются в ноль ни при каких i. Если допустить, что прогоночные коэффициенты находятся точно, а в y_N допущена ошибка ε_N, то погрешность решения ε_i будет удовлетворять однородному уравнению $\varepsilon_i = \alpha_i \varepsilon_{i+1}$, то есть не будет нарастать при выполнении условия $|\alpha_i| \leqslant 1$ для всех i. В этом случае можно говорить об устойчивости метода прогонки.

Теорема 2.1 *Если коэффициенты системы (2.1) удовлетворяют условиям* $|b_0| \geqslant 0$, $|c_0| > 0$, $|a_N| \geqslant 0$, $|c_N| > 0$,

$$|a_i| > 0, \quad |b_i| > 0, \quad |c_i| \geqslant |a_i| + |b_i|, \quad i = 1, 2, ..., N-1, \tag{2.2}$$

$$|c_0| \geqslant |b_0|, \quad |c_N| \geqslant |a_N|, \tag{2.3}$$

причем хотя бы в одном из неравенств (2.2) или (2.3) выполняется строгое неравенство, то есть для матрицы системы (2.1) имеет место диагональное преобладание, то $c_i - a_i \alpha_{i-1} \neq 0$ *и* $|\alpha_i| \leqslant 1$ *для всех* $i = 0, 1, ..., N-1$.

Доказательство. Из условий теоремы следует, что $0 \leqslant |\alpha_0| = \dfrac{|b_0|}{|c_0|} \leqslant 1$. Пусть $|\alpha_i| \leqslant 1$, тогда

$$|\alpha_{i+1}| = \frac{|b_{i+1}|}{|c_{i+1} - a_{i+1}\alpha_i|} \leqslant \frac{|b_{i+1}|}{|c_{i+1}| - |a_{i+1}|} \leqslant \frac{|b_{i+1}|}{|b_{i+1}|} = 1.$$

Следовательно, $|\alpha_i| \leqslant 1$ для всех $i = 0, 1, ..., N-1$. Кроме того, имеют место неравенства

$$|c_i - a_i \alpha_{i-1}| \geqslant |c_i| - |a_i| \cdot |\alpha_{i-1}| \geqslant |b_i| + |a_i| \cdot (1 - |\alpha_{i-1}|) \geqslant |b_i| > 0, \ \ i \leqslant N-1,$$

откуда получаем, что $c_i - a_i \alpha_{i-1} \neq 0$ при $i \leqslant N-1$.

Остается показать, что $c_N - a_N\alpha_{N-1} \neq 0$. По условию хотя бы в одном из неравенств (2.2) или (2.3) выполняется строгое неравенство. Если $|c_N| > |a_N|$, то $c_N - a_N\alpha_{N-1} \neq 0$, так как $|\alpha_{N-1}| \leqslant 1$. Если существует $1 \leqslant i_0 \leqslant N-1$, такое что $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$, то $|c_{i_0} - a_{i_0}\alpha_{i_0-1}| > |b_{i_0}|$, откуда следует, что $|\alpha_{i_0}| < 1$. Тогда по индукции получаем, что $|\alpha_i| < 1$ для всех $i \geqslant i_0 + 1$. Следовательно, $|c_N - a_N\alpha_{N-1}| > 0$, так как $|\alpha_{N-1}| < 1$. Наконец, если $|c_0| > |b_0|$, то неравенство $|\alpha_i| < 1$ выполняется, начиная с $i = 0$, а значит, как и в предыдущем случае, $|c_N - a_N\alpha_{N-1}| > 0$.

Сформулированные условия являются лишь *достаточными* условиями корректности метода монотонной правой прогонки. Их можно ослабить, разрешив некоторым из коэффициентов a_i и b_i обращаться в ноль.

«Частично» пользуясь методом прогонки, можно упростить решение системы с разреженной матрицей, фрагмент которой является трехдиагональным. На этой идее основаны различные методы окаймления. В качестве примера рассмотрим систему с «почти» трехдиагональной матрицей:

$$\begin{cases} a_1 y_N - c_1 y_1 + b_1 y_2 = -f_1, \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \;\; i = 2, 3, ..., N-1, \\ a_N y_{N-1} - c_N y_N + b_N y_1 = -f_N. \end{cases} \tag{2.4}$$

Такая алгебраическая система возникает при отыскании *периодического* решения системы трехточечных уравнений:

$$y_{i+N} = y_i$$

при условии, что:

$$a_{i+N} = a_i, \quad b_{i+N} = b_i, \quad c_{i+N} = c_i, \quad f_{i+N} = f_i.$$

Запишем систему (2.4) в матричном виде $A_N Y_N = -F_N$, где $Y_N = (y_1, ..., y_N)^{\mathrm{T}}$, $F_N = (f_1, ..., f_N)^{\mathrm{T}}$, а матрица A_N имеет вид:

$$A_N = \begin{bmatrix} -c_1 & b_1 & 0 & 0 & \dots & 0 & 0 & a_1 \\ a_2 & -c_2 & b_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & a_3 & -c_3 & b_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -c_{N-2} & b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & a_{N-1} & -c_{N-1} & b_{N-1} \\ b_N & 0 & 0 & 0 & \dots & 0 & a_N & -c_N \end{bmatrix}. \tag{2.5}$$

Присутствие ненулевых элементов в правом верхнем и левом нижнем углах матрицы (2.5) не позволяет решать систему (2.4) обычным методом прогонки. Запишем систему (2.4) в виде:

$$\begin{cases} A_{N-1}Y_{N-1} + U_{N-1}y_N = -F_{N-1}, \\ V_{N-1}Y_{N-1} - c_N y_N = -f_N, \end{cases} \tag{2.6}$$

где $Y_{N-1} = (y_1, ..., y_{N-1})^{\mathrm{T}}$, $F_{N-1} = (f_1, ..., f_{N-1})^{\mathrm{T}}$,

$$A_{N-1} = \begin{bmatrix} -c_1 & b_1 & 0 & 0 & \dots & 0 & 0 \\ a_2 & -c_2 & b_2 & 0 & \dots & 0 & 0 \\ 0 & a_3 & -c_3 & b_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -c_{N-2} & b_{N-2} \\ 0 & 0 & 0 & 0 & \dots & a_{N-1} & -c_{N-1} \end{bmatrix}, \quad U_{N-1} = \begin{bmatrix} a_1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ b_{N-1} \end{bmatrix},$$

$$V_{N-1} = \begin{bmatrix} b_N & 0 & 0 & 0 & \dots & 0 & a_N \end{bmatrix}.$$

Решение первого блока уравнений системы (2.6) будем искать в виде:

$$Y_{N-1} = P_{N-1} + y_N Q_{N-1},$$

где $P_{N-1} = (p_1, p_2, ..., p_{N-1})^{\mathrm{T}}$ и $Q_{N-1} = (q_1, q_2, ..., q_{N-1})^{\mathrm{T}}$ — решения задач:

$$A_{N-1}P_{N-1} = -F_{N-1}, \quad A_{N-1}Q_{N-1} = -U_{N-1}. \tag{2.7}$$

Поскольку A_{N-1} — трехдиагональная матрица, то системы (2.7) решаются обычной прогонкой:

$$\begin{cases} \alpha_i = \dfrac{b_i}{c_i - \alpha_{i-1}a_i}, \quad \beta_i = \dfrac{f_i + a_i\beta_{i-1}}{c_i - \alpha_{i-1}a_i}, \quad \gamma_i = \dfrac{a_i\gamma_{i-1}}{c_i - \alpha_{i-1}a_i}, \quad i = 2, 3, ..., N; \\ \alpha_1 = \dfrac{b_1}{c_1}, \quad \beta_1 = \dfrac{f_1}{c_1}, \quad \gamma_1 = \dfrac{a_1}{c_1}; \end{cases} \tag{2.8}$$

$$\begin{cases} p_{N-1} = \beta_{N-1}, \quad q_{N-1} = \alpha_{N-1} + \gamma_{N-1}; \\ p_i = \alpha_i p_{i+1} + \beta_i, \quad q_i = \alpha_i q_{i+1} + \gamma_i, \quad i = N-2, N-3, ..., 1. \end{cases} \tag{2.9}$$

Выразим теперь y_N:

$$V_{N-1}Y_{N-1} - c_N y_N = V_{N-1}P_{N-1} + y_N V_{N-1}Q_{N-1} - c_N y_N = -f_N \quad \Rightarrow$$

$$y_N = \frac{f_N + b_N p_1 + a_N p_{N-1}}{c_N - b_N q_1 - a_N q_{N-1}} = \frac{f_N + b_N p_1 + a_N \beta_{N-1}}{c_N - b_N q_1 - a_N(\alpha_{N-1} + \gamma_{N-1})} \cdot \frac{c_N - a_N\alpha_{N-1}}{c_N - a_N\alpha_{N-1}} =$$

$$= \frac{\beta_N + \alpha_N p_1}{1 - \alpha_N q_1 - \gamma_N}.$$

Теперь, зная y_N и коэффициенты p_i, q_i, $i = 1, ..., N-1$, находим решение системы:

$$y_i = p_i + y_N q_i, \quad i = 1, 2, ..., N-1.$$

Изложенный алгоритм требует $Q = 14N - 8$ действий.

Теорема 2.2 *Пусть коэффициенты системы (2.4) удовлетворяют условиям:*

$$|a_i| > 0, \ \ |b_i| > 0, \ \ |c_i| \geqslant |a_i| + |b_i|, \ \ i = 1, 2, ..., N, \tag{2.10}$$

и существует такое число $1 \leqslant i_0 \leqslant (N-1)$, *что* $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$. *Тогда* $c_i - a_i\alpha_{i-1} \neq 0$, $|\alpha_i| \leqslant 1$ *и* $|\alpha_i| + |\gamma_i| \leqslant 1$ *для всех* $i = 2, 3, ..., N$, *и выполняется условие* $1 - \alpha_N q_1 - \gamma_N \neq 0$. Доказательство. Если выполнены условия (2.10), то из теоремы 2.1 следует, что $c_i - a_i\alpha_{i-1} \neq 0$ и $|\alpha_i| \leqslant 1$ для всех возможных i.

Так как $|a_1| + |b_1| \leqslant |c_1|$, то $|\alpha_1| + |\gamma_1| \leqslant 1$. Предположим, что $|\alpha_i| + |\gamma_i| \leqslant 1$. Тогда

$$|\alpha_{i+1}| + |\gamma_{i+1}| = \frac{|b_{i+1}| + |a_{i+1}| \cdot |\gamma_i|}{|c_{i+1} - a_{i+1}\alpha_i|} \leqslant \frac{|a_{i+1}| + |b_{i+1}| - |a_{i+1}|(1 - |\gamma_i|)}{|c_{i+1}| - |a_{i+1}| \cdot |\alpha_i|} \leqslant$$

$$\leqslant \frac{|a_{i+1}| + |b_{i+1}| - |a_{i+1}| \cdot |\alpha_i|}{|c_{i+1}| - |a_{i+1}| \cdot |\alpha_i|} = 1.$$

Следовательно, условие $|\alpha_i| + |\gamma_i| \leqslant 1$ выполнено для всех i.

Так как $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$, то $|\alpha_{i_0}| + |\gamma_{i_0}| < 1$, а значит по индукции $|\alpha_i| + |\gamma_i| < 1$ для всех $i \geqslant i_0$. В том числе, $|\alpha_{N-1}| + |\gamma_{N-1}| < 1$, то есть $|q_{N-1}| \leqslant |\alpha_{N-1}| + |\gamma_{N-1}| < 1$. Но тогда по индукции получаем, что $|q_i| \leqslant 1$ для всех возможных i. В частности, $|q_1| \leqslant 1$, и

$$|1 - \alpha_N q_1 - \gamma_N| \geqslant 1 - |\alpha_N| - |\gamma_N| > 0.$$

2.2 Метод немонотонной прогонки

Рассмотрим систему с трехдиагональной матрицей

$$\begin{cases} -c_0 y_0 + b_0 y_1 = -f_0, \ \ (i = 0), \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \ \ (i = 1, 2, ..., N-1), \\ a_N y_{N-1} - c_N y_N = -f_N, \ \ (i = N), \end{cases}$$

для которой не выполнены условия диагонального преобладания. При формальном использовании метода Гаусса без выбора главного элемента на l-м шаге приведения исходной

системы к системе с верхнетреугольной матрицей получаем:

$$\begin{cases} -(c_l - a_l\alpha_{l-1})y_l + b_l y_{l+1} = -(f_l + a_l\beta_{l-1}), \ \ (i = l), \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \ \ (l+1 \leqslant i \leqslant N-1), \\ a_N y_{N-1} - c_N y_N = -f_N, \ \ (i = N). \end{cases} \tag{2.11}$$

Если матрица системы имеет диагональное преобладание, то $c_l - a_l\alpha_{l-1} \neq 0$, то есть первое уравнение системы (2.11) можно представить в виде

$$y_l = \alpha_l y_{l+1} + \beta_l, \quad \alpha_l = \frac{b_l}{c_l - a_l\alpha_{l-1}}, \ \ \beta_l = \frac{f_l + a_l\beta_{l-1}}{c_l - a_l\alpha_{l-1}},$$

причем $|\alpha_l| \leqslant 1$. Если же диагональное преобладание не имеет места, то нельзя гарантировать, что $c_l - a_l\alpha_{l-1} \neq 0$, и даже если это так, условие $|\alpha_l| \leqslant 1$ может быть не выполнено. В этом случае метод монотонной прогонки уже не применим.

Предположим, что исходная система имеет единственное решение. Тогда, применяя метод Гаусса с выбором главного элемента в строке, на l-м шаге получаем:

$$\begin{cases} -C_l y_{m_l} + b_l y_{l+1} = -F_l, \ \ (i = l), \\ A_l y_{m_l} - c_{l+1} y_{l+1} + b_{l+1} y_{l+2} = -\Phi_l, \ \ (i = l+1), \\ a_{l+2} y_{l+1} - c_{l+2} y_{l+2} + b_{l+2} y_{l+3} = -f_{l+2}, \ \ (i = l+2), \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \ \ (l+3 \leqslant i \leqslant N-1), \\ a_N y_{N-1} - c_N y_N = -f_N, \ \ (i = N). \end{cases} \tag{2.12}$$

Если $l = 0$, то имеют место равенства $C_0 = c_0$, $A_0 = a_1$, $F_0 = f_0$, $\Phi_0 = f_1$ и $m_0 = 0$. При дальнейшем преобразовании системы (2.12) возможно два варианта. В первом случае $|C_l| \geqslant |b_l|$. Тогда первое уравнение системы (2.12) можно переписать в виде

$$y_{m_l} - \alpha_l y_{l+1} = \beta_l, \quad \alpha_l = \frac{b_l}{C_l}, \ \ \beta_l = \frac{F_l}{C_l},$$

где $|\alpha_l| \leqslant 1$, или же в виде $y_{\theta_{l+1}} = \alpha_l y_{m_{l+1}} + \beta_l$, где $\theta_{l+1} = m_l$, $m_{l+1} = l+1$.

Исключая из уравнения, соответствующего $i = l+1$, слагаемое $A_l y_{m_l}$, получаем

$$\begin{cases} -C_{l+1} y_{m_{l+1}} + b_{l+1} y_{l+2} = -F_{l+1}, \ \ (i = l+1), \\ A_{l+1} y_{m_{l+1}} - c_{l+2} y_{l+2} + b_{l+2} y_{l+3} = -\Phi_{l+1}, \ \ (i = l+2), \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \ \ (l+3 \leqslant i \leqslant N-1), \\ a_N y_{N-1} - c_N y_N = -f_N, \ \ (i = N), \end{cases} \tag{2.13}$$

где $m_{l+1} = l+1$, $C_{l+1} = c_{l+1} - A_l\alpha_l$, $F_{l+1} = \Phi_l + A_l\beta_l$, $A_{l+1} = a_{l+2}$, $\Phi_{l+1} = f_{l+2}$.

Во втором случае $|C_l| < |b_l|$. При этом первое уравнение системы (2.12) можно переписать в виде

$$y_{l+1} - \alpha_l y_{m_l} = \beta_l, \quad \alpha_l = \frac{C_l}{b_l}, \ \ \beta_l = -\frac{F_l}{b_l},$$

где опять же $|\alpha_l| \leqslant 1$, или, что то же самое, в виде $y_{\theta_{l+1}} = \alpha_l y_{m_{l+1}} + \beta_l$, где $\theta_{l+1} = l+1$, $m_{l+1} = m_l$.

Полученное уравнение используем для исключения y_{l+1} из уравнения, соответствующего $i = l+1$. В результате снова придем к системе (2.13), но теперь $m_{l+1} = m_l$, $C_{l+1} = c_{l+1}\alpha_l - A_l$, $F_{l+1} = \Phi_l - c_{l+1}\beta_l$, $A_{l+1} = a_{l+2}\alpha_l$, $\Phi_{l+1} = f_{l+2} + a_{l+2}\beta_l$.

Итак, *алгоритм немонотонной прогонки* следующий:

1) $C = c_0$, $A = a_1$, $F = f_0$, $\Phi = f_1$ и $m_0 = 0$;

2) для всех $i = 0, 1, ..., N-1$ выполняются действия (прямой ход)

а) если $|C| \geqslant |b_i|$, то $\alpha_i = \dfrac{b_i}{C}$, $\beta_i = \dfrac{F}{C}$, $C = c_{i+1} - A\alpha_i$, $F = \Phi + A\beta_i$, $\theta_{i+1} = m_i$, $m_{i+1} = i+1$, для $i \leqslant N-2$ находим $A = a_{i+2}$, $\Phi = f_{i+2}$;

б) если $|C| < |b_i|$, то $\alpha_i = \dfrac{C}{b_i}$, $\beta_i = -\dfrac{F}{b_i}$, $C = c_{i+1}\alpha_i - A$, $F = \Phi - c_{i+1}\beta_i$, $\theta_{i+1} = i+1$, $m_{i+1} = m_i$, для $i \leqslant N-2$ находим $A = a_{i+2}\alpha_i$, $\Phi = f_{i+2} + a_{i+2}\beta_i$;

3) вычисляем $y_{m_N} = \dfrac{F}{C}$;

4) для всех $i = N-1$, $N-2$, ..., 0 совершаем обратный ход, находя

$$y_{\theta_{i+1}} = \alpha_i y_{m_{i+1}} + \beta_i.$$

Если исходная система не вырождена, то коэффициенты C_l и b_l одновременно в ноль обратиться не могут. Это обеспечивает корректность процесса. Так как по построению для всех α_i справедливы неравенства $|\alpha_i| \leqslant 1$. Следовательно, описанный процесс устойчив по отношению к ошибкам округления. В самом худшем случае, когда вычисления все время ведутся по сценарию (б), общее число действий составляет $Q = 12N$.

2.3 Метод матричной прогонки

Пусть система сеточных уравнений, полученных в результате разностной аппроксимации задачи, может быть записана в виде

$$\begin{cases} -C_0\mathbf{Y}_0 + B_0\mathbf{Y}_1 = -\mathbf{F}_0, \quad (i=0), \\ A_i\mathbf{Y}_{i-1} - C_i\mathbf{Y}_i + B_i\mathbf{Y}_{i+1} = -\mathbf{F}_i, \quad (1 \leqslant i \leqslant N-1), \\ A_N\mathbf{Y}_{N-1} - C_N\mathbf{Y}_N = -F_N, \quad (i=N), \end{cases} \tag{2.14}$$

где $\mathbf{Y}_i$ — неизвестные векторы размерности M_i, $\mathbf{F}_i$ — заданные векторы размерности M_i, C_i — квадратные матрицы размера $M_i \times M_i$, A_i и B_i — прямоугольные матрицы размеров $M_i \times M_{i-1}$ и $M_i \times M_{i+1}$ соответственно.

Для решения системы (2.14) рассмотрим метод матричной прогонки. По аналогии с системой скалярных трехточечных уравнений будем искать решение в виде

$$\mathbf{Y}_i = \alpha_i \mathbf{Y}_{i+1} + \beta_i, \quad i = N-1,\ N-2,\ ...,\ 1,\ 0,$$

где α_i — матрицы размера $M_i \times M_{i+1}$, β_i — векторы размерности M_i. При этом формально получаем:

$$\alpha_0 = C_0^{-1} B_0, \quad \alpha_i = (C_i - A_i \alpha_{i-1})^{-1} B_i, \ \ i = 1, 2, ..., N-1;$$

$$\beta_0 = C_0^{-1} \mathbf{F}_0, \quad \beta_i = (C_i - A_i \alpha_{i-1})^{-1} (\mathbf{F}_i + A_i \beta_{i-1}), \ \ i = 1, 2, ..., N;$$

$$\mathbf{Y}_N = \beta_N.$$

Если матрицы C_0 и $(C_i - A_i \alpha_{i-1})$ для $i = 1, 2, ..., N$ не вырождены, то алгоритм матричной прогонки корректен. Будем говорить, что алгоритм устойчив, если $\|\alpha_i\| \leqslant 1$ для $i = 1, 2, ..., N$.

Теорема 2.3 *Если C_i для $i = 0, 1, ..., N$ — невырожденные матрицы, а A_i и B_i — ненулевые матрицы для $i = 1, 2, ..., N-1$ и выполнены условия $\|C_0^{-1} B_0\| \leqslant 1$, $\|C_N^{-1} A_N\| \leqslant 1$, $\|C_i^{-1} A_i\| + \|C_i^{-1} B_i\| \leqslant 1$ для $i = 1, 2, ..., N-1$, причем хотя бы в одном из неравенств имеет место строгое неравенство, то алгоритм метода матричной прогонки устойчив и корректен.*

Если все матрицы A_i, B_i, C_i квадратные и имеют размер $M \times M$, а все векторы $\mathbf{Y}_i$ и $\mathbf{F}_i$ имеют размерность M, то для решения системы (2.14) методом матричной прогонки потребуется $Q = O(M^3 N + M^2 N)$ действий.

3 Итерационные методы

3.1 Итерационные методы решения разностных уравнений как задачи на установление

Пусть требуется решить уравнение

$$Au = f, \tag{3.1}$$

где A — линейный самосопряженный положительно определенный оператор, действующий в вещественном гильбертовом пространстве H со скалярным произведением (y, v) и нормой $\|y\| = \sqrt{(y, y)}$, $f \in H$ — произвольная функция.

Уравнению (3.1) можно поставить в соответствие абстрактную задачу Коши:

$$\begin{cases} \dfrac{dv}{dt} + Av = f, \ \ t > 0, \\ v(0) = v_0, \end{cases} \tag{3.2}$$

где v_0 — произвольный элемент пространства H, $v(t)$ — функция со значениями в H.

Покажем, что при сформулированных условиях на оператор A решение задачи (3.2) стремится по норме к решению задачи (3.1):

$$\lim_{t\to\infty} \|v(t) - y\| = 0.$$

Для этого введем функцию $z(t) = v(t) - y$. Она будет удовлетворять задаче Коши с однородным уравнением:

$$\begin{cases} \dfrac{dz}{dt} + Az = 0, \ \ t > 0, \\ z(0) = v_0 - y. \end{cases} \tag{3.3}$$

Умножая уравнение в задаче (3.3) скалярно на $z(t)$, получаем:

$$\left(\frac{dz}{dt}, z\right) + (Az, z) = 0.$$

Так как

$$\left(\frac{dz}{dt}, z\right) = \frac{1}{2}\frac{d}{dt}(z, z) = \frac{1}{2}\frac{d}{dt}\|z(t)\|^2,$$

и по условию существует такое $\delta > 0$, что $(Az, z) \geqslant \delta\|z\|^2$, то приходим к неравенству

$$\frac{d}{dt}\|z(t)\|^2 + 2\delta\|z\|^2 \leqslant 0,$$

из которого следует, что

$$e^{2\delta t}\frac{d}{dt}\|z(t)\|^2 + 2\delta e^{2\delta t}\|z\|^2 = \frac{d}{dt}\left(e^{2\delta t}\|z\|^2\right) \leqslant 0.$$

Интегрируя последнее неравенство, получаем:

$$e^{2\delta t}\|z(t)\|^2 \leqslant \|z(0)\|^2,$$

откуда находим

$$\|v(t) - y\| = \|z(t)\| \leqslant e^{-\delta t}\|v_0 - y\| \to 0 \ \text{ при } \ t \to \infty.$$

Следовательно, для того чтобы найти приближенное решение задачи (3.1), можно построить разностную схему для задачи (3.2) и вычислять ее решение до тех пор, пока не будет выполнено условие

$$\left\|\frac{dv}{dt}\right\| < \varepsilon.$$

Начальное условие в задаче (3.2) выбирается произвольно.

Если Λ — линейный самосопряженный положительно определенный разностный оператор, аппроксимирующий оператор A, то разностную схему для задачи (3.2) можно записать в виде

$$\begin{cases} B_k \dfrac{y^{k+1} - y^k}{\tau_{k+1}} + \Lambda y^k = f, \quad k = 0, 1, 2, ... \\ y^0 = v_0, \end{cases} \tag{3.4}$$

где τ_{k+1} — шаги в общем случае неравномерной сетки по времени, B_k — обратимый оператор.

Разностную схему (3.4) можно интерпретировать как итерационный процесс:

$$y^0 = v_0, \quad B_k y^{k+1} = (B_k - \tau_{k+1}\Lambda) y^k + \tau_{k+1} f, \quad k = 0, 1, 2, ...$$

для нахождения приближенного решения сеточного уравнения

$$\Lambda y = f. \tag{3.5}$$

В качестве практического критерия прекращения итераций целесообразно выбирать условие малости невязки:

$$\|\Lambda y - f\| < \varepsilon,$$

где ε — требуемая точность.

В данном случае мы получили двухслойную итерационную схему, или, что то же самое, одношаговый итерационный метод, так как для вычисления y^{k+1} используется только предыдущая итерация y^k.

В общем случае говорят, что итерационный процесс сходится, если $\|y^k - y\| \to 0$ при $k \to \infty$. Вычисления прекращаются, если достигнуто условие $\|\Lambda y^k - f\| < \varepsilon$.

Различие между итерационными схемами и схемами для нестационарных задач заключается в следующем:

1) итерационная схема (3.4) точно аппроксимирует исходное стационарное сеточное уравнение (3.5), то есть точное решение y уравнения (3.5) удовлетворяет уравнению (3.4) при любых B_k и τ_{k+1} ;

2) выбор параметров τ_{k+1} и операторов B_k следует подчинить лишь требованиям сходимости итерационного процесса и экономичности, то есть минимума арифметических операций для получения решения исходной задачи с заданной точностью, тогда как в случае нестационарной задачи выбор шага подчинен, прежде всего, требованию аппроксимации.

Итерационную схему (то есть набор $\{\tau_{k+1}\}$ и $\{B_k\}$) следует выбирать так, чтобы минимальное число $n(\varepsilon)$ итераций, при котором достигается заданная точность ε, сочеталось

с минимальным числом действий Q_k для нахождения k-й итерации. Тогда минимальным будет общее число арифметических операций $Q(\varepsilon)$, которое нужно выполнить, чтобы получить при помощи метода (3.4) решение уравнения (3.5) с заданной точностью $\varepsilon > 0$ при любом выборе начального приближения:

$$Q(\varepsilon) = \sum_{k=1}^{n(\varepsilon)} Q_k.$$

3.2 Итерационный метод переменных направлений

В качестве первого примера рассмотрим разностную задачу Дирихле для уравнения Пуассона в прямоугольнике $\bar{G} = \{x = (x^1, x^2),\ x^p \in [0, l_p],\ p = 1, 2\}$:

$$\begin{cases} \Lambda y = -f(x), \ \ x \in \omega_h, \\ y = \mu(x), \ \ x \in \gamma_h, \end{cases} \tag{3.6}$$

где $\Lambda = \Lambda_1 + \Lambda_2$, $\Lambda_p y = y_{\bar{x}^p x^p}$, $p = 1, 2$,

$$\omega_h + \gamma_h = \left\{(x_n^1, x_m^2),\ x_n^1 = nh_1,\ x_m^2 = mh_2, n = 0, ..., N,\ m = 0, ..., M,\ Nh_1 = l_1,\ Mh_2 = l_2\right\}.$$

Для нахождения приближенного решения задачи (3.6) используем итерационную схему переменных направлений:

$$\begin{cases} \dfrac{y^{j+\frac{1}{2}} - y^j}{\tau_{j+1}^{(1)}} = \Lambda_1 y^{j+\frac{1}{2}} + \Lambda_2 y^j + f(x), \ \ x \in \omega_h \\ \\ y^{j+\frac{1}{2}} = \mu(x), \ \ x \in \gamma_h, \\ \dfrac{y^{j+1} - y^{j+\frac{1}{2}}}{\tau_{j+1}^{(2)}} = \Lambda_1 y^{j+\frac{1}{2}} + \Lambda_2 y^{j+1} + f(x), \ \ x \in \omega_h \\ \\ y^{j+1} = \mu(x), \ \ x \in \gamma_h, \end{cases} \tag{3.7}$$

где $j = 0, 1, ..., y^0 = u_0(x)$ — начальное приближение, которое по возможности нужно выбирать удовлетворяющим граничным условиям задачи, $\tau_{j+1}^{(1)} > 0$ и $\tau_{j+1}^{(2)} > 0$ — итерационные параметры, подлежащие выбору исходя из условия минимума числа итераций.

Погрешность $z^{j+1} = y^{j+1} - y$ удовлетворяет задаче:

$$\begin{cases} \dfrac{z^{j+\frac{1}{2}} - z^j}{\tau_{j+1}^{(1)}} = \Lambda_1 z^{j+\frac{1}{2}} + \Lambda_2 z^j, \;\; x \in \omega_h \\ \\ z^{j+\frac{1}{2}} = 0, \;\; x \in \gamma_h, \\ \dfrac{z^{j+1} - z^{j+\frac{1}{2}}}{\tau_{j+1}^{(2)}} = \Lambda_1 z^{j+\frac{1}{2}} + \Lambda_2 z^{j+1}, \;\; x \in \omega_h \\ \\ z^{j+1} = 0, \;\; x \in \gamma_h, \\ z^0 = y^0 - y. \end{cases}$$

Если ввести пространство H_h сеточных функций со скалярным произведением

$$(y, v) = \sum_{x \in \omega_h} y(x) v(x) h_1 h_2,$$

все элементы которого ограничены по норме $\|y\| = \sqrt{(y, y)}$ и обращаются в нуль на γ_h, то операторы $A_p y = -\Lambda_p y$, $p = 1, 2$, будут самосопряженными, положительно определенными и перестановочными, причем

$$\lambda_p^{\min} E \leqslant A_p \leqslant \lambda_p^{\max} E, \;\; p = 1, 2 \tag{3.8}$$

где

$$\lambda_p^{\min} = \frac{4}{h_p^2} \sin^2 \frac{\pi h_p}{2 l_p}, \quad \lambda_p^{\max} = \frac{4}{h_p^2} \cos^2 \frac{\pi h_p}{2 l_p}, \;\; p = 1, 2.$$

Задачу для погрешности z^{j+1} можно переписать в виде:

$$\begin{cases} \left(E + \tau_{j+1}^{(1)} A_1\right) z^{j+\frac{1}{2}} = \left(E - \tau_{j+1}^{(1)} A_2\right) z^j, \\ \left(E + \tau_{j+1}^{(2)} A_2\right) z^{j+1} = \left(E - \tau_{j+1}^{(2)} A_1\right) z^{j+\frac{1}{2}}, \end{cases}$$

где $z^0 = y^0 - y$, $j = 0, 1, ...$, или, исключая $z^{j+\frac{1}{2}}$ и пользуясь перестановочностью операторов A_1 и A_2, в виде $z^{j+1} = S_{j+1} z^j$, где $S_{j+1} = S_{j+1}^{(1)} S_{j+1}^{(2)}$ — оператор перехода со слоя на слой, а операторы $S_{j+1}^{(1)}$ и $S_{j+1}^{(2)}$ имеют вид:

$$S_{j+1}^{(1)} = \left(E + \tau_{j+1}^{(1)} A_1\right)^{-1} \left(E - \tau_{j+1}^{(2)} A_1\right), \quad S_{j+1}^{(2)} = \left(E + \tau_{j+1}^{(2)} A_2\right)^{-1} \left(E - \tau_{j+1}^{(1)} A_2\right).$$

Таким образом,

$$z^n = T_n z^0, \tag{3.9}$$

где $T_n = \prod_{j=1}^{n} S_j$ — разрешающий оператор, причем $T_n^* = T_n$.

Итерационные параметры $\tau_{j+1}^{(1)}$ и $\tau_{j+1}^{(2)}$ подбираются таким образом, чтобы для получения точности ε затратить минимальное число шагов. Для этого необходимо точно знать границы спектра оператора A. Выбор оптимальных (по Жордану) параметров обеспечивает минимум $\|T_n\|$.

3.3 Выбор оптимальных параметров итерационной схемы переменных направлений (по Жордану)

Из равенства (3.9) следует оценка

$$\|z^n\| \leqslant \|T_n\| \cdot \|z^0\|.$$

Величина $\|T_n\|$ зависит от параметров $\tau_j^{(1)}$ и $\tau_j^{(2)}$, $j = 1, 2, ...$ Задача состоит в отыскании таких параметров $\tau_1^{(1)}$, $\tau_2^{(1)}$, ..., $\tau_n^{(1)}$ и $\tau_1^{(2)}$, $\tau_2^{(2)}$, ..., $\tau_n^{(2)}$, где число итераций $n = n(\varepsilon)$, необходимое для достижения точности ε, задано:

$$\min_{\{\tau_j^{(1)}, \tau_j^{(2)}\}} \|T_n\| = q_n.$$

Из оценок (3.8) следует, что спектр оператора A_p принадлежит отрезку:

$$\delta_p \leqslant \lambda(A_p) \leqslant \Delta_p, \;\; p = 1, 2.$$

Заменим операторы A_1 и A_2 операторами A_1' и A_2', спектры которых совпадают:

$$\eta E \leqslant A_p' \leqslant E, \;\; p = 1, 2, \;\; \eta > 0.$$

Положим $A_1 = (qE - rA_1')^{-1}(A_1' - pE)$, $A_2 = (qE + rA_2')^{-1}(A_2' + pE)$, где r, p, q — числа, подлежащие выбору, и введем параметры:

$$\omega^{(1)} = \frac{\tau^{(1)} - r}{q - \tau^{(1)} p}, \;\; \omega^{(2)} = \frac{\tau^{(2)} + r}{q + \tau^{(2)} p}.$$

При этом получим:

$$S_j = \tilde{S}_j^{(1)} \tilde{S}_j^{(2)},$$

где

$$\tilde{S}_j^{(1)} = (E + \omega_j^{(1)} A_1')^{-1}(E - \omega_j^{(2)} A_1'), \qquad \tilde{S}_j^{(2)} = (E + \omega_j^{(2)} A_2')^{-1}(E - \omega_j^{(1)} A_2').$$

Выразим $\|T_n\|$ через собственные значения операторов A_1' и A_2':

$$\alpha_{k_1} = \lambda_{k_1}^{(1)}(A_1'), \qquad \beta_{k_2} = \lambda_{k_2}^{(2)}(A_2'), \;\; k_\alpha = 1, 2, ..., N_\alpha, \;\; \alpha = 1, 2.$$

Так как $A_1' A_2' = A_2' A_1'$, то они имеют общую систему собственных функций, то же, что и операторы A_1, A_2, A и T_n. Пусть $\lambda_k(T_n)$ — собственные значения оператора T_n. Поскольку

$$\lambda(\tilde{S}^{(1)}) = \frac{1 - \omega^{(2)}\alpha}{1 + \omega^{(1)}\alpha}, \qquad \lambda(\tilde{S}^{(2)}) = \frac{1 - \omega^{(1)}\beta}{1 + \omega^{(2)}\beta},$$

то

$$\lambda(T_n) = \prod_{j=1}^{n} \frac{1-\omega_j^{(2)}\alpha}{1+\omega_j^{(1)}\alpha} \cdot \frac{1-\omega_j^{(1)}\beta}{1+\omega_j^{(2)}\beta}, \tag{3.10}$$

причем

$$0 < \eta \leqslant \alpha_{k_1} \leqslant 1; \quad 0 < \eta \leqslant \beta_{k_2} \leqslant 1; \quad k_\alpha = 1, 2, ..., N_\alpha; \ \alpha = 1, 2.$$

Норма оператора T_n равна наибольшему собственному значению этого оператора:

$$\|T_n\| = \max_k \lambda_k(T_n).$$

Заменим в выражении (3.10) α_{k_1} и β_{k_2} непрерывно меняющимися аргументами α и β. При этом максимум правой части (3.10), вообще говоря, увеличивается, то есть

$$\|T_n\| \leqslant \max_{\alpha,\beta\in[\eta,1]} \left| \prod_{j=1}^{n} \frac{1-\omega_j^{(2)}\alpha}{1+\omega_j^{(1)}\alpha} \cdot \frac{1-\omega_j^{(1)}\beta}{1+\omega_j^{(2)}\beta} \right|. \tag{3.11}$$

Поскольку α и β меняются на отрезке $[\eta, 1]$, а $\omega_j^{(1)}$ и $\omega_j^{(2)}$ входят в формулу (3.11) симметрично, то можно положить $\alpha = \beta$, $\omega_j^{(1)} = \omega_j^{(2)} = \omega_j$ (минимум произведения будет достигаться, когда каждый сомножитель будет достигать минимума). В результате приходим к следующей задаче: требуется найти параметры ω_1, ω_2, ..., ω_n, при которых достигается минимум

$$\min_{\{\omega_j\}} \max_{\alpha\in[\eta,1]} \prod_{j=1}^{n} \left(\frac{1-\omega_j\alpha}{1+\omega_j\alpha} \right)^2. \tag{3.12}$$

Задача (3.12) называется задачей минимакса, ее решение известно. Приведем окончательные формулы для вычисления оптимальных параметров $\tau_j^{(1)}$, $\tau_j^{(2)}$. Постоянные p, q, r, η находятся из условий: $\alpha = \beta = \eta$ при $\lambda(A_1) = \delta_1$, $\lambda(A_2) = \delta_2$ и $\alpha = \beta = 1$ при $\lambda(A_1) = \Delta_1$, $\lambda(A_2) = \Delta_2$, и выражаются формулами:

$$\begin{gathered} t = \sqrt{\frac{(\Delta_1-\delta_1)(\Delta_2-\delta_2)}{(\Delta_1+\delta_1)(\Delta_2+\delta_2)}}, \quad \eta = \frac{1-t}{1+t}, \quad \varkappa = \frac{(\Delta_1-\delta_1)\Delta_2}{(\Delta_2+\delta_1)\Delta_1}, \\ p = \frac{\varkappa - t}{\varkappa + t}, \quad r = \frac{\Delta_1 - \Delta_2 + (\Delta_1+\Delta_2)p}{2\Delta_1\Delta_2}, \quad q = r + \frac{1-p}{\Delta_1}, \end{gathered} \tag{3.13}$$

причем $\varkappa > t$, $p > 0$.

Пусть задана точность $\varepsilon > 0$ итерационного процесса и известны границы δ_α, Δ_α операторов A_α. По формулам (3.13) находим η, p, q и r. После этого можно определить число итераций $n(\varepsilon)$, обеспечивающих заданную точность $\varepsilon > 0$: $\|y_n - u\| \leqslant \varepsilon\|y_0 - u\|$. Справедлива приближенная формула:

$$n(\varepsilon) \approx \frac{1}{\pi^2} \ln\frac{4}{\varepsilon} \ln\frac{4}{\eta}. \tag{3.14}$$

Введем обозначения:

$$\theta = \frac{\eta^2}{16}\left(1 + \frac{\eta^2}{2}\right), \quad \sigma = \frac{2j-1}{2n}, \ \ j = 1, 2, ..., n.$$

Для определения ω_j имеет место формула:

$$\omega_j = \frac{(1+2\theta)(1+\theta^\sigma)}{2\theta^{\sigma/2}(1+\theta^{1-\sigma}+\theta^{1+\sigma})}, \;\; j = 1, 2, ..., n \tag{3.15}$$

Остается определить $\tau_j^{(1)}$ и $\tau_j^{(2)}$, исходя из формул

$$\omega = \frac{\tau^{(1)} - r}{q - \tau^{(1)} p}, \qquad \omega = \frac{\tau^{(2)} + r}{q + \tau^{(2)} p}. \tag{3.16}$$

В частном случае, когда $\delta_1 = \delta_2 = \delta$ и $\Delta_1 = \Delta_2 = \Delta$ имеем:

$$p = r = 0, \;\; q = \frac{1}{\Delta}, \;\; \eta = \frac{\delta}{\Delta}, \;\; \varkappa = \frac{1-\eta}{1+\eta}.$$

Преобразование операторов при этом принимает вид:

$$A_1 = \Delta A_1', \quad A_2 = \Delta A_2', \quad \omega^{(1)} = \Delta\tau^{(1)}, \quad \omega^{(2)} = \Delta\tau^{(2)}.$$

Условие $\omega^{(1)} = \omega^{(2)}$ дает $\tau^{(1)} = \tau^{(2)} = \tau$.

3.4 Итерационный метод на основе эволюционно-факторизованной схемы

Рассмотрим краевую задачу для эллиптического уравнения без смешанных производных в прямоугольном параллелепипеде G (прямоугольнике в двумерном случае) либо области G, составленной из параллелепипедов:

$$\begin{cases} Lu = -f(x), \; x \in G, \;\; L = \sum\limits_{p=1}^{3} \dfrac{\partial}{\partial x_p}\left(k_p(x)\dfrac{\partial}{\partial x_p}\right), \; k_p(x) > 0, \\ u|_{\partial G} = \mu(x). \end{cases} \tag{3.17}$$

Поставим в области G задачу на установление

$$\begin{cases} \dfrac{\partial v}{\partial t} = Lv + f(x), \; x \in G, \;\; t > 0, \\ v|_{\partial G} = \mu(x), \quad v|_{t=0} = v_0(x), \end{cases} \tag{3.18}$$

где $v_0(x)$ — произвольная функция, желательно, удовлетворяющая граничным условиям задачи.

Введем в расчетной области пространственную сетку $\overline{\omega_h} = \omega_h + \gamma_h$, где ω_h — множество внутренних узлов, а γ_h — множество граничных узлов, а также сетку по времени с шагом τ, и построим для задачи (3.18) эволюционно-факторизованную схему:

$$\underbrace{(E - 0.5\tau\Lambda_1)\,(E - 0.5\tau\Lambda_2)\,\underbrace{(E - 0.5\tau\Lambda_3)y_t}_{y_2}}_{y_1} = \Lambda y + f,$$

где Λ_p — разностная аппроксимация L_p, $p = 1, 2, 3$, $\Lambda = \sum_{p=1}^{3} \Lambda_p$.

Переход между целыми слоями по времени осуществляется в три шага:

$$\begin{cases} (E - 0.5\tau\Lambda_1)y_1 = \Lambda y + f, \\ y_1|_{\gamma_h^1} = (E - 0.5\tau\Lambda_2)y_2|_{\gamma_h^1} = 0, \end{cases}$$

$$\begin{cases} (E - 0.5\tau\Lambda_2)y_2 = y_1, \\ y_2|_{\gamma_h^2} = (E - 0.5\tau\Lambda_3)\mu_t|_{\gamma_h^2} = 0, \end{cases}$$

$$\begin{cases} (E - 0.5\tau\Lambda_3)y_t = y_2, \\ y|_{\gamma_h} = \mu, \end{cases}$$

где γ_h^1 — граничные узлы по направлению x^1, γ_h^2 — граничные узлы по направлению x^2.

Пусть $\lambda_p^{(i)}$ — собственные значения соответствующих задач Штурма-Лиувилля:

$$\begin{cases} \Lambda_p y + \lambda_p y = 0, \;\; x \in \omega_h, \;\; p = 1, 2, 3, \\ y = 0, \;\; x \in \gamma_h^p. \end{cases}$$

Тогда множители роста ρ для рассматриваемой эволюционно-факторизованной схемы удовлетворяют уравнению:

$$(1 + 0.5\tau\lambda_1)(1 + 0.5\tau\lambda_2)(1 + 0.5\tau\lambda_3)(\rho - 1) = -\tau(\lambda_1 + \lambda_2 + \lambda_3).$$

Одномерный случай:

$$\rho = \frac{1 - 0.5\tau\lambda_1}{1 + 0.5\tau\lambda_1}.$$

Двумерный случай:

$$\rho = \frac{1 - 0.5\tau\lambda_1}{1 + 0.5\tau\lambda_1} \cdot \frac{1 - 0.5\tau\lambda_2}{1 + 0.5\tau\lambda_2}.$$

Трехмерный случай:

$$\rho = 1 - \frac{\tau(\lambda_1 + \lambda_2 + \lambda_3)}{(1 + 0.5\tau\lambda_1)(1 + 0.5\tau\lambda_2)(1 + 0.5\tau\lambda_3)}.$$

Счет на установление по эволюционно-факторизованной схеме сходится значительно быстрее, если вместо постоянного шага по времени τ (итерационного параметра) использовать специальный набор τ_s, $s = 1, 2, ..., S$. В настоящий момент наиболее эффективным является так называемый логарифмический набор τ_s. Идея его построения состоит в том, чтобы на каждом шаге по времени гасить одну какую-либо гармонику в численном решении или группу соседних гармоник (то есть добиваться минимума по модулю соответствующих множителей роста).

В одномерном случае целесообразно выбирать

$$\tau_{\min} = \frac{2}{\lambda^{\max}}, \quad \tau_{\max} = \frac{2}{\lambda^{\min}},$$

так как при этом полностью гасятся первая и последняя гармоники, а остальные подавляются частично.

Из тех же соображений в двумерном случае выбирают

$$\tau_{\min} = \frac{2}{\max\left\{\lambda_1^{\max}, \lambda_2^{\max}\right\}}, \quad \tau_{\max} = \frac{2}{\min\left\{\lambda_1^{\min}, \lambda_2^{\min}\right\}}.$$

В трехмерном случае уравнение $\rho(\tau) = 0$ либо не имеет положительных корней, либо имеет три вещественных корня, два из которых положительны, а один отрицателен. Если положительных корней у уравнения $\rho(\tau) = 0$ нет, то $\tau_{\min}$ соответствует минимуму $\rho(\tau)$, если положить $\lambda_p = \lambda_p^{\max}$:

$$\tau_{\min} = \tau_*, \quad \frac{1}{\tau *} = -\sqrt{\frac{b}{3}}\cos\left(\varphi + \frac{2\pi}{3}\right),$$

$$b = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3, \qquad c = \lambda_1\lambda_2\lambda_3, \qquad \varphi = \frac{1}{3}\arccos\left(-c\left(\frac{b}{3}\right)^{-3/2}\right),$$

если $\rho(\tau_*) \geqslant 0$. Если же $\rho(\tau_*) < 0$, то $\tau_{\min}$ — минимальный положительный корень $\rho(\tau)$, который может быть найден по формуле:

$$\frac{2}{\tau_\pm} = \frac{a}{3} - 2\sqrt{\frac{f}{3}}\cos\left(\theta \mp \frac{2\pi}{3}\right), \quad \tau_{\min} = \tau_-,$$

где

$$a = \lambda_1 + \lambda_2 + \lambda_3, \qquad f = \frac{(\lambda_1 - \lambda_2)^2}{2} + \frac{(\lambda_1 - \lambda_3)^2}{2} + \frac{(\lambda_2 - \lambda_3)^2}{2},$$

$$g = \frac{\lambda_1^2(-2\lambda_1 + 3\lambda_2 + 3\lambda_3)}{27} + \frac{\lambda_2^2(-2\lambda_2 + 3\lambda_1 + 3\lambda_3)}{27} + \frac{\lambda_3^2(-2\lambda_3 + 3\lambda_2 + 3\lambda_1)}{27} + \frac{14\lambda_1\lambda_2\lambda_3}{9},$$

$$\theta = \frac{1}{3}\arccos\left[\frac{g}{2}\left(\frac{f}{3}\right)^{-3/2}\right].$$

Аналогично вычисляется $\tau_{\max}$. При этом во всех выражениях $\lambda_p^{\max}$ необходимо заменить на $\lambda_p^{\min}$. Если выполняется условие $\rho(\tau_*) \geqslant 0$, то $\tau_{\max} = \tau_*$. Если $\rho(\tau_*) < 0$, то $\tau_{\max} = \tau_+$.

Общий вид логарифмического набора итерационных параметров следующий:

$$\ln \tau_s = \frac{1}{2}\ln(\tau_{\min}\tau_{\max}) + \frac{1}{2}\ln\left(\frac{\tau_{\max}}{\tau_{\min}}\right) f(s), \quad s = 0, 1, ..., S,$$

где $f(s) \in [-1, 1]$ — порождающая функция, которая должна быть монотонной и нечетной.

Равномерный набор: $f_{\text{р}} = \dfrac{2s}{S} - 1$.

Чебышевский набор: $f_{\text{ч}} = -\dfrac{\pi s}{S}$.

Наилучшими свойствами в плане скорости сходимости обладает линейная комбинация равномерного и чебышевского наборов:

$$f(s) = C \cdot f_{\text{р}}(s) + (1 - C) \cdot f_{\text{ч}}(s), \tag{3.19}$$

где рекомендуемое значение $C = \dfrac{\pi}{\pi + 2}$.

На практике точные границы спектра оператора Λ, как правило, неизвестны. Для них можно получить только более или менее точные оценки. Поэтому в расчетах вместо границ спектра рекомендуется брать величины $\dfrac{\lambda_{\min}}{q}$ и $\lambda_{\max} q$, где значения $\lambda_{\min}$ и $\lambda_{\max}$ получены оценочно, параметр $q > 1$, то есть производить расширение границ спектра. Практическая рекомендация: $q \sim 3$. При этом будет иметь место некоторая потеря скорости расчетов, но для набора (3.19) она не очень велика.

3.5 Некоторые итерационные методы, не сводящиеся к задаче на установление

Существует много актуальных задач, к которым не применим счет на установление с помощью эволюционно-факторизованной схемы. Например, это задачи в криволинейных областях, для которых приходится использовать треугольную сетку, задачи со смешанными производными и несамосопряженными операторами, задачи для уравнений высоких порядков и т.д. При их разностной аппроксимации возникает необходимость решать СЛАУ вида:

$$Au = f,$$

где A — сильно разреженная матрица размера $M \times M$, u, f — векторы размера $M \times 1$.

Для задач рассматриваемого класса наиболее быстро сходятся многошаговые итерационные методы сопряженных направлений. Они сводятся к минимизации некоторой квадратичной формы $\Phi(u)$ и построению соответствующего ортонормированного базиса. Теоретически, они позволяют найти точное решение системы, но на практике в силу того, что M достаточно велико, расчеты прекращают по достижении определенной точности. Критерием прекращения итераций выступает малость невязки.

3.5.1 Метод сопряженных градиентов

Метод применим для эрмитовых знакоопределенных матриц A. При этом осуществляется поиск экстремума квадратичной формы

$$\Phi(u) = (u, Au - 2f).$$

В качестве начального приближения выбирается произвольное u_1. Вспомогательные величины: r_s — невязка, p_s — направление очередного спуска, q_s — вспомогательный вектор. Итерационный алгоритм имеет вид:

$$r_s = \begin{cases} Au_s - f, \ \ s = 1, \\ r_{s-1} - \dfrac{q_{s-1}}{(q_{s-1}, p_{s-1})}, \ \ s = 2, 3, ... \end{cases} \qquad p_s = p_{s-1} + \frac{r_s}{(r_s, r_s)}, \ \ p_0 = 0,$$
$$q_s = Ap_s, \quad u_{s+1} = u_s - \frac{p_s}{(q_s, p_s)}.$$

Условие прекращения итераций: $\|r_s\| < \varepsilon$. При этом выполняется неравенство

$$\|u_s - u_{\text{точн}}\| \leqslant \varepsilon \frac{\lambda_{\max}}{\lambda_{\min}},$$

то есть для хорошо обусловленных матриц метод дает приемлемые результаты.

Для знакопеременных матриц сходимость метода не доказана, для неэрмитовых матриц метод не сходится.

3.5.2 Метод сопряженных невязок

Метод применим для эрмитовых знаконеопределенных матриц. Начальное приближение u_1 выбирается произвольно. Итерационный алгоритм имеет вид:

$$r_s = \begin{cases} Au_s - f, \ \ s = 1, \\ r_{s-1} - q_{s-1}\dfrac{(r_{s-1}, q_{s-1})}{(q_{s-1}, q_{s-1})}, \ \ s = 2, 3, ... \end{cases} \qquad g_s = Ar_s, \quad p_s = p_{s-1} + \frac{r_s}{(r_s, g_s)}, \ \ p_0 = 0,$$
$$q_s = \begin{cases} Ap_s, \ \ s = 1, \\ q_{s-1} + \dfrac{g_s}{(r_s, g_s)}, \ \ s = 2, 3, ... \end{cases}$$
$$u_{s+1} = u_s - p_s \frac{(r_s, q_s)}{(q_s, q_s)}.$$

Для неэрмитовых матриц метод не сходится.

3.5.3 Метод Крейга

Метод рассчитан на произвольные (даже не квадратные) матрицы A. В случае неквадратной матрицы метод сходится к псевдорешению системы. Начальное приближение u_1 выбирается произвольно, $p_0 = 0$. Итерационный алгоритм имеет вид:

$$r_s = Au_s - f, \ \ p_s = p_{s-1} + \frac{r_s}{(r_s, r_s)}, \ \ q_s = A^{\text{H}} p_s, \ \ u_{s+1} = u_s - \frac{q_s}{(q_s, q_s)}.$$

Данный метод целесообразно применять лишь для несимметричных матриц.

4 Контрольные вопросы

1) Сформулируйте и докажите теорему о достаточных условиях корректности и устойчивости метода монотонной правой прогонки.

2) В чем состоит идея метода окаймления? Проиллюстрируйте этот метод на примере системы с "почти трехдиагональной"матрицей для периодической сеточной функции. 3) Сформулируйте алгоритм немонотонной прогонки.

4) Сформулируйте алгоритм матричной прогонки и достаточные условия его применимости.

5) В чем состоит идея итерационных методов решения систем сеточных уравнений? Как эти методы связаны с задачами на установление? Запишите каноническую форму двухслойной итерационной схемы.

6) Изложите итерационный метод переменных направлений для приближенного решения задачи для уравнения Пуассона в прямоугольнике.

7) Составьте эволюционно-факторизованную итерационную схему для уравнения Пуассона в прямоугольнике и прямоугольном параллелепипеде. Как выбираются оптимальные итерационные параметры?